

# APPROXIMATELY OPTIMAL SCHEDULING OF AN M/G/1 QUEUE WITH HEAVY TAILS

VIJAY KAMBLE, JEAN WALRAND,\* *University of California, Berkeley*

## Abstract

Distributions with a heavy tail are difficult to estimate. If the design of a scheduling policy is sensitive to the details of heavy tail distributions of the service times, an approximately optimal solution is difficult to obtain. This paper shows that the optimal scheduling of an M/G/1 queue with heavy tailed service times does not present this difficulty and that an approximately optimal strategy can be derived by truncating the distributions.

*Keywords:* Scheduling; Heavy Tails; M/G/1 Queue; Approximation

2010 Mathematics Subject Classification: Primary 60K25

Secondary 90B15; 60G17

## 1. Introduction

We consider the scheduling of jobs in an M/G/1 queue when the service time distribution  $F$  of the jobs may have a heavy tail, but has a finite mean and variance. At any given time, the server remembers how long he has worked on each job in the queue and chooses which job to work on, possibly in a preemptive way. The objective of the server is to find a scheduling policy that minimizes the mean sojourn time of the jobs in the system. Can one derive a good scheduling policy by neglecting the details of the tail of  $F$ ? Let  $X$  be a service time of a typical job and let the rate of arrival be  $\lambda$ . The result of the paper is that there exist constants  $K_1 > 0$  and  $K_2 > 0$  dependent only on  $E(X)$ ,  $E(X^2)$  and  $\lambda$  such that if  $s > 0$  is such that

$$K_1 E[X \mathbf{1}_{\{X > s\}}] + K_2 E[X^2 \mathbf{1}_{\{X > s\}}] \leq \epsilon$$

---

\* Postal address: Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA, 94720, USA

\* Email address: {vjk, wlr}@eecs.berkeley.edu

for  $\epsilon > 0$ , then one can derive an  $\epsilon$ -optimal scheduling policy by truncating the distribution of  $X$  to  $s$ . Finding the value of  $s$  that corresponds to a given  $\epsilon > 0$  is simpler than estimating the details of the tail of the distribution of  $X$ , so this result has practical significance. Specifically, this  $\epsilon$ -optimal policy agrees with the optimal policy for the distribution  $F$  truncated at  $s$ , except that it switches to an arbitrarily chosen work-conserving policy when it discovers a job whose service time exceeds  $s$ . The same result holds when there are no arrivals, which we call the static case.

## 2. The static case

There are  $N$  jobs whose service times are independent and distributed according to distribution  $F$ . Fix a threshold  $s > 0$ . Let  $F_s$  be the distribution  $F$  truncated at  $s$ , i.e.  $F_s(a) = \frac{F(a \wedge s)}{F(s)}$  for  $a \in \mathbb{R}$ . The admissible scheduling policies are preemptive and such that the server bases his decision on which job to serve only on the amount of time he has spent on the different jobs so far and on which jobs are still in the system. That is, the server does not know the residual service times of the unfinished jobs. A policy is optimal if it minimizes the sum of the expected job completion times. Let  $V^*$  be the expected sum of the completion times of the jobs under the optimal scheduling policy  $u$ . Let  $u_s$  be the optimal scheduling policy for the system with the truncated service time distribution  $F_s$ . Also, let  $\tilde{u}_s$  be the policy that agrees with  $u_s$  until it discovers a job with a service time larger than  $s$ , after which it switches to a given arbitrarily chosen work conserving strategy  $\tilde{u}$ . Finally, let  $V^s$  be the expected sum of the completion times under policy  $\tilde{u}_s$ .

**Theorem 2.1.** (Static Case.) *Assume that  $F$  has a finite mean. Then,*

$$\limsup_{s \rightarrow \infty} V^s - V^* = 0.$$

This result says that the scheduling policy for the truncated distribution is almost optimal for the original distribution. The policy for the truncated distribution switches to any arbitrary work-conserving policy if it discovers a job with a service time larger than the truncation threshold. The intuitive justification for the result is that, on the event that no job has a service time larger than  $s$ , the policy  $\tilde{u}_s$  makes the optimal decisions. In the rare event that a job has a service time larger than  $s$ , the policy

incurs an extra cost but its expected value goes to zero as  $s$  increases to infinity.

*Proof.* Let the service times be the random variables  $X_1, \dots, X_N$  and let  $C(u)$  be the random variable denoting the sum of the completion times under a fixed policy  $u$ . Consider the event  $A = \{X_i \leq s, i = 1, \dots, N\}$ . We have

$$\begin{aligned} V^* &= E[C(u^*)] = E[C(u^*) | A]P(A) + E[C(u^*) | A^c]P(A^c) \\ &\geq E[C(u_s) | A]P(A) + E[C(u^*) | A^c]P(A^c) \\ &\geq E[C(u_s) | A]P(A). \end{aligned} \tag{1}$$

The first inequality holds since on the event that service times of all jobs are less than the threshold  $s$ , the optimal policy for the truncated service times achieves a lower cost than the optimal policy  $u^*$ . Now, let us turn to the analysis of the cost under policy  $\tilde{u}_s$ . We have

$$\begin{aligned} V^s &= E[C(\tilde{u}_s)] = E[C(\tilde{u}_s) | A]P(A) + E[C(\tilde{u}_s) | A^c]P(A^c) \\ &= E[C(u_s) | A]P(A) + E[C(\tilde{u}_s) | A^c]P(A^c). \end{aligned} \tag{2}$$

Thus, from (1) and (2) we have that

$$V^s - V^* \leq E[C(\tilde{u}_s) | A^c]P(A^c). \tag{3}$$

Now suppose that, if at least one of the jobs has a service time which exceeds the threshold, we are told the exact realizations of  $X_1, \dots, X_N$ . With this additional information the worst sum of completion times is achieved by the policy where the server finishes serving all the jobs together. In that case, the completion time of each job is the sum of the file sizes. The expected cost under this policy is an upper bound for the cost achieved by any work-conserving policy which works without this additional information of the service times. Moreover, the upper bound for the cost under this policy is when the service times of all the files exceed the threshold  $s$ . Thus we have

$$\begin{aligned} E[C(\tilde{u}_s) | A^c]P(A^c) &\leq N^2 E[X_i | X_i > s]P(A^c) \\ &\leq N^3 E[X_i | X_i > s]P(X_i > s) \end{aligned} \tag{4}$$

where the last inequality is a result of the union bound. Now, since  $E(X_i) < \infty$ , it follows that

$$\lim_{s \rightarrow \infty} E[X_i \mathbf{1}_{\{X_i \geq s\}}] = 0.$$

This completes the proof. ■

### 3. The dynamic case

We now consider the optimal scheduling problem of an M/G/1 queue. Since the process describing the number of jobs in the system is a renewal process that does not depend on the work-conserving policy, we consider a single busy period and our aim is to find a work conserving scheduling strategy  $u$  which minimizes the expected value of the sum of sojourn times of all jobs in the busy period. Formally, for a busy period labelled  $B$ , we want to minimize  $E(C(u))$  where

$$C(u) = \sum_{i=1}^{N_B} S_i \quad (5)$$

where  $N_B$  is a random variable denoting the number of jobs that arrive in the busy period  $B$ , and  $S_i$  is a random variable denoting the sojourn time of job  $i$  when policy  $u$  is used to schedule the jobs. Let  $u^*$  be the optimal policy with service times distributed according to  $F$ , and  $V^*$  be the corresponding optimal expected cost. Fix a threshold  $s > 0$  and let  $u_s$  be the optimal policy when all service times are distributed according to the truncated distribution  $F_s$ . Also, let  $\tilde{u}_s$  be the policy that uses the policy  $u_s$  till it discovers a job whose service time exceeds  $s$ , after which it switches to an arbitrary but work conserving policy  $\tilde{u}$ . Let  $V^s$  be the corresponding expected average of sojourn times of the jobs in a busy period under this policy.

**Theorem 3.1.** (Dynamic Case.) *Assume that distribution of the service time of a job  $F$  has a finite mean and variance. Let  $X$  be the service time of a typical job. Then, there exist constants  $K_1$  and  $K_2$  dependent only on the mean and variance of  $F$ , such that*

$$V^s - V^* \leq K_1 E[X \mathbf{1}_{\{X > s\}}] + K_2 E[X^2 \mathbf{1}_{\{X > s\}}].$$

Hence

$$\limsup_{s \rightarrow \infty} V^s - V^* = 0.$$

To prove the theorem, we will first need the following lemma:

**Lemma 1.** *Let  $X$  be the service time of a typical job. Consider a busy period  $B$ . Let  $W$  be the duration of the busy period and  $N_B$  be the number of jobs that arrive during*

the busy period. For a threshold  $s$ , let  $A$  be the event that the service time of every job in the busy period is less than the threshold. Then

1.  $E[W \mid A^c] \leq \frac{E[X|X>s]}{D^2}$ .
2.  $E[N_B \mid A^c] \leq \frac{1}{D} + \frac{\lambda E[X|X>s]}{D^2}$ .

where  $D = 1 - \lambda E(X)$ .

*Proof.* We first compute an upper bound for  $P(A^c)$ . This is the probability that at least one job in the busy period has a service time that exceeds the threshold. Suppose that the policy is first come, first served except that the lowest priority is given to the first job that arrives in the busy period. The number of jobs in the system as a function of time in a typical busy period under this scheduling policy is illustrated in fig. 1. Let

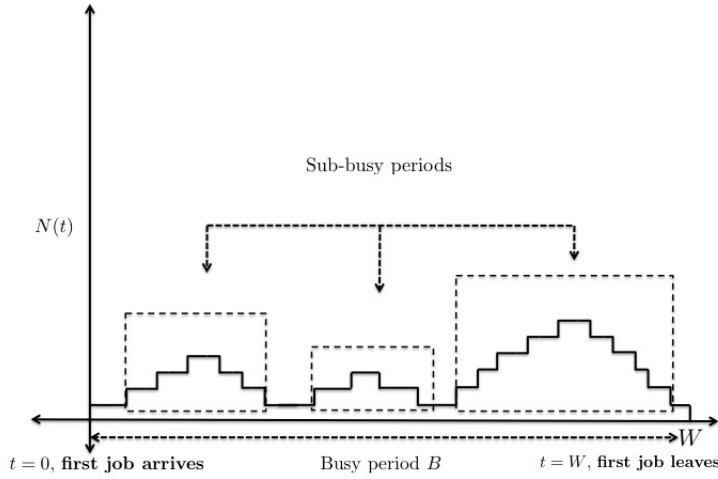


FIGURE 1: The number of jobs in the system as a function of time in a busy period  $B$  under the policy which is first come first served except that lowest priority is given to the first job that arrives. “Sub-busy periods” interrupt the service of the first job until it finishes.

$X$  denote the service time of the first job. Then we have the following expression:

$$P(A^c) = P(X > s) + P(X \leq s)E[1 - \exp(-\lambda P(A^c)X) \mid X \leq s].$$

This holds because while working on the first job with service time  $X$ , new busy periods with at least one job with service time that exceeds the threshold  $s$  and busy periods with no job with service time exceeding  $s$  arrive as independent Poisson processes with rate  $\lambda P(A^c)$  and  $\lambda P(A)$  respectively. These “sub-busy periods” interrupt the service of the first job until it finishes. But since  $1 - \exp(-x) \leq x$ , we have

$$P(A^c) \leq P(X > s) + P(X \leq s)E[\lambda P(A^c)X \mid X \leq s].$$

Thus we have

$$P(A^c) \leq \frac{P(X > s)}{1 - \lambda P(X \leq s)E[X \mid X \leq s]}.$$

But since  $P(X \leq s)E[X \mid X \leq s] \leq E(X)$ , we have

$$P(A^c) \leq \frac{P(X > s)}{1 - \lambda E(X)}. \quad (6)$$

Now to prove part (1), we find an upper bound for  $E[W \mid A^c]$ . We again assume that the policy is first come, first served except that the lowest priority is given to the first job that arrives in the busy period. We use a first step argument that decomposes  $W$  into  $X$  plus the duration of the busy periods that arrive while the server processes the first job. Assume that  $K(X)$  busy periods arrive during the processing of the first job. Among these  $K(X)$  sub-busy periods,  $N(X)$  have at least one job with service time that exceeds  $s$  and  $M(X) = K(X) - N(X)$  have all jobs with service times less than  $s$ . Note that as mentioned before, conditional on  $X$ , the random variables  $N(X)$  and  $M(X)$  are Poisson with mean  $\lambda X P(A^c)$  and  $\lambda X P(A)$ , respectively. Define  $\alpha := E[W \mid A^c]$ ,  $\delta = E[W \mid A]$  and  $\gamma = P[X > s \mid A^c] = \frac{P(X > s)}{P(A^c)}$ . Then,

$$\alpha = E[W \mid X > s, A^c]\gamma + E[W \mid X \leq s, A^c](1 - \gamma).$$

Now,

$$E[W \mid X > s, A^c] = E[W \mid X > s] = E[X \mid X > s] + \lambda E[X \mid X > s]E(W).$$

Also,

$$\begin{aligned} E[W \mid X \leq s, A^c] &= E[X + \alpha E[N(X) \mid N(X) > 0, X] \mid X \leq s, N(X) > 0] \\ &\quad + E[\delta E[M(X) \mid N(X) > 0, X] \mid X \leq s, N(X) > 0]. \end{aligned}$$

Furthermore,

$$E[N(X) \mid N(X) > 0, X] = \frac{\lambda P(A^c)X}{1 - \exp(-\lambda P(A^c)X)}.$$

Since  $\frac{x}{1+\exp(-x)} \leq 1+x$ , we have

$$E[N(X) \mid N(X) > 0, X] \leq 1 + \lambda P(A^c)X. \quad (7)$$

Also  $M(X)$  and  $N(X)$  are independent conditional on  $X$ . Hence,

$$\begin{aligned} \alpha &\leq \gamma (E[X \mid X > s] + \lambda E[X \mid X > s]E(W)) \\ &\quad + (1 - \gamma) (E[(1 + \lambda P(A^c)X)\alpha + X + \lambda P(A)X\delta \mid X \leq s, N(X) > 0]) \\ &= \gamma(1 + \lambda E(W))E[X \mid X > s] \\ &\quad + (1 - \gamma)(\alpha + (1 + \lambda E(W))E[X \mid X \leq s, N(X) > 0]). \end{aligned}$$

Now denote  $E[X \mid X \leq s, N(X) > 0] = \phi(s)$ . Then we have,

$$\alpha \leq \gamma^{-1}(1 + \lambda E(W))(\gamma E[X \mid X > s] + (1 - \gamma)\phi(s)). \quad (8)$$

Now from the the definition of  $\gamma$  we have

$$\begin{aligned} E[W \mid A^c] &\leq \frac{P(A^c)}{P(X > s)}(1 + \lambda E(W)) \left( \frac{P(X > s)}{P(A^c)} E[X \mid X > s] + \left(1 - \frac{P(X > s)}{P(A^c)}\right) \phi(s) \right). \end{aligned}$$

Now, (a) from the upper bound on  $P(A^c)$  in (6), (b) since  $E(W) = \frac{E(X)}{1 - \lambda E(X)}$  and (c) since  $\phi(s) = E[X \mid X \leq s, N(X) > 0] \leq E[X \mid X > s]$ , we finally have

$$E[W \mid A^c] \leq C \quad (9)$$

where

$$\begin{aligned} C &= \frac{1}{D^2} E[X \mid X > s] \text{ and} \\ D &= 1 - \lambda E(X). \end{aligned}$$

We thus have the required result.

Now to prove part (2), we find an upper bound for  $E[N_B \mid A^c]$ . Again define  $\beta := E[N_B \mid A^c]$ ,  $\rho = E[N_B \mid A]$  and  $\gamma = P[X > s \mid A^c] = \frac{P(X > s)}{P(A^c)}$ . Then,

$$\beta = E[N_B \mid X > s, A^c]\gamma + E[N_B \mid X \leq s, A^c](1 - \gamma).$$

Now,

$$E[N_B \mid X > s, A^c] = E[N_B \mid X > s] = 1 + \lambda E[X \mid X > s] E(N_B).$$

Also,

$$\begin{aligned} E[N_B \mid X \leq s, A^c] &= E[1 + \beta E[N(X) \mid N(X) > 0, X] \mid X \leq s, N(X) > 0] \\ &\quad + E[\rho E[M(X) \mid N(X) > 0, X] \mid X \leq s, N(X) > 0]. \end{aligned}$$

Hence by same arguments as in the case before,

$$\begin{aligned} \beta &\leq \gamma (1 + \lambda E[X \mid X > s] E(N_B)) \\ &\quad + (1 - \gamma) (E[(1 + \lambda P(A^c)X)\beta + 1 + \lambda P(A)X\rho \mid X \leq s, N(X) > 0]) \\ &= \gamma (1 + \lambda E[X \mid X > s] E(N_B)) + (1 - \gamma) (\beta + 1 + \lambda E(N_B)\phi(s)). \end{aligned}$$

Therefore,

$$\beta \leq \gamma^{-1} \left[ 1 + \lambda E(N_B) (\gamma E[X \mid X > s] + (1 - \gamma)\phi(s)) \right]. \quad (10)$$

Now from the the definition of  $\gamma$  we have

$$\begin{aligned} E[N_B \mid A^c] &\leq \frac{P(A^c)}{P(X > s)} \left[ 1 + \lambda E(N_B) \left( \frac{P(X > s)}{P(A^c)} E[X \mid X > s] + \left( 1 - \frac{P(X > s)}{P(A^c)} \right) \phi(s) \right) \right]. \end{aligned}$$

Now again (a) from the upper bound on  $P(A^c)$  in (6), (b) since  $E(N_B) = \frac{1}{1 - \lambda E(X)}$  and (c) since  $\phi(s) = E[X \mid X \leq s, N(X) > 0] \leq E[X \mid X > s]$ , we finally have

$$E[N_B \mid A^c] \leq P \quad (11)$$

where

$$\begin{aligned} P &= \frac{1}{D} + \frac{\lambda}{D^2} E[X \mid X > s] \text{ and} \\ D &= 1 - \lambda E(X) \end{aligned}$$

which is the required result. ■



We now prove the main theorem.

*Proof.* The system starts empty and we consider the first busy period. Let  $X$  be the service time of the first job that arrives. Let  $A$  be the event that every service time in the busy period is less than the threshold  $s$ . Then we have

$$\begin{aligned} V^* &= E[C(u^*)] = E[C(u^*) | A]P(A) + E[C(u^*) | A^c]P(A^c) \\ &\geq E[C(u^*) | A]P(A) \geq E[C(u_s) | A]P(A). \end{aligned}$$

The first inequality holds because we ignore the second term in the first expression while the second inequality holds because, on the event  $A$ , policy  $u_s$  gives a lower cost than policy  $u^*$ . By definition of  $u_s$  and  $\tilde{u}_s$ , one has

$$V^s = E[C(u_s) | A]P(A) + E[C(\tilde{u}_s) | A^c]P(A^c).$$

We thus have

$$V^s - V^* \leq E[C(\tilde{u}_s) | A^c]P(A^c).$$

Under any work conserving policy, the sum of sojourn times of the jobs in a busy period, is upper bounded by the total duration of the busy period times the number of files that arrived in the busy period, and this random variable is independent of the scheduling policy. Let  $W$  be the random variable denoting the total duration of the busy period and  $N_B$  is the number of files that arrived in the busy period. Then we have

$$V^s - V^* \leq E[WN_B | A^c]P(A^c). \quad (12)$$

The rest of the proof is providing an upper bound on this quantity. We already have an upper bound for  $P(A^c)$  from (6). We move on to find an upper bound for  $E[WN_B | A^c]$ . Again we use a technique similar to the one used in proving the lemma. Say that the policy is first come, first served except that the first job to arrive in the busy period is given the lowest priority. We use a first step argument that decomposes  $WN_B$  into the contribution to this quantity by this first job in the busy period plus the contribution from all the sub-busy periods that arrive and interrupt while the server processes the first job. As in the proof of the lemma, assume that  $K(X)$  sub-busy periods arrive during the processing of the first job. Among these  $K(X)$  busy periods,  $N(X)$  have at least one job with service time that exceeds  $s$  and  $M(X) = K(X) - N(X)$  have all jobs

with service times less than  $s$ . Here as before, given  $X$ , the random variables  $N(X)$  and  $M(X)$  are independent poisson with mean  $\lambda X P(A^c)$  and  $\lambda X P(A)$  respectively. We define two sets of notation for the durations of these busy periods. The duration of all the busy periods are denoted by  $\{W_k; k = 1 \dots, K(X)\}$  and the number of jobs in each of these busy periods is denoted by  $\{N_{B_k}; k = 1 \dots, K(X)\}$ . We will find it convenient to also denote the busy periods with at least one job with service time exceeding threshold  $s$  by  $\{\hat{W}_i; i = 1, \dots, N(X)\}$  and the number of jobs in these busy periods by  $\{N_{\hat{B}_i}; i = 1, \dots, N(X)\}$ . Similarly, the busy periods with no job with service time exceeding threshold  $s$  are denoted by  $\{\tilde{W}_j; j = 1, \dots, M(X)\}$  and the number of jobs in these busy periods are denoted by  $\{N_{\tilde{B}_j}; j = 1, \dots, M(X)\}$ . Also define  $\gamma = P[X > s \mid A^c] = \frac{P(X > s)}{P(A^c)}$ . Then,

$$E[WN_B \mid A^c] = E[WN_B \mid X > s, A^c]\gamma + E[WN_B \mid X \leq s, A^c](1 - \gamma).$$

We now have the following decomposition:

$$\begin{aligned} WN_B &= (X + \sum_{j=1}^{K(X)} W_j)(1 + \sum_{j=1}^{K(X)} N_{B_j}) \\ &= X + \sum_{i=1}^{K(X)} W_i + \sum_{j=1}^{K(X)} W_i N_{B_i} + \sum_{k=1}^{K(X)} N_{B_k} (X + \sum_{j \neq k; j=1}^{K(X)} W_j) \\ &= X + \sum_{i=1}^{N(X)} \hat{W}_i + \sum_{j=1}^{M(X)} \tilde{W}_j + \sum_{i=1}^{N(X)} \hat{W}_i N_{\hat{B}_i} + \sum_{j=1}^{M(X)} \tilde{W}_j N_{\tilde{B}_j} \\ &\quad + \sum_{i=1}^{N(X)} N_{\hat{B}_i} (X + \sum_{l \neq i; l=1}^{N(X)} \hat{W}_l + \sum_{j=1}^{M(X)} \tilde{W}_j) \\ &\quad + \sum_{j=1}^{M(X)} N_{\tilde{B}_j} (X + \sum_{i=1}^{N(X)} \hat{W}_i + \sum_{m \neq j; m=1}^{M(X)} \tilde{W}_m) \end{aligned} \tag{13}$$

Now,

$$\begin{aligned} E[WN_B \mid X > s, A^c] &= E[WN_B \mid X > s] \\ &= E[X + \sum_{i=1}^{K(X)} W_i + \sum_{j=1}^{K(X)} W_i N_{B_i} \mid X > s] \\ &\quad + E[\sum_{k=1}^{K(X)} N_{B_k} (X + \sum_{j \neq k; j=1}^{K(X)} W_j) \mid X > s]. \end{aligned}$$

Now for each sub-busy period  $i$ ,  $N_{\hat{B}_i}$  is independent of  $X$  and  $\sum_{j \neq k; j=1}^{K(X)} W_j$ . We thus have

$$\begin{aligned} E\left[\sum_{k=1}^{K(X)} N_{B_k}(X + \sum_{j \neq k; j=1}^{K(X)} W_j) \mid X\right] &= E\left[\sum_{k=1}^{K(X)} N_{B_k} E[X + \sum_{j \neq k; j=1}^{K(X)} W_j \mid X] \mid X\right] \\ &= \lambda X E(N_B)(X + (\lambda X - 1)E(W)). \end{aligned}$$

Thus we have

$$\begin{aligned} E[WN_B \mid X > s, A^c] &= E[X \mid X > s] [1 + \lambda E(W) + \lambda E(WN_B)] \\ &\quad + E[\lambda X E(N_B)(X + (\lambda X - 1)E(W)) \mid X > s] \\ &\leq E[X \mid X > s] (1 + \lambda E(W) + \lambda E(WN_B)) \\ &\quad + E[X^2 \mid X > s] \lambda E(N_B)(1 + \lambda E(W)) \end{aligned} \quad (14)$$

where the inequality arises since we ignore the negative term. Also,

$$\begin{aligned} E[WN_B \mid X \leq s, A^c] &= E[WN_B \mid X \leq s, N(X) > 0] \\ &= E[X + \sum_{i=1}^{N(X)} \hat{W}_i + \sum_{j=1}^{M(X)} \tilde{W}_j \mid X \leq s, N(X) > 0] \\ &\quad + E\left[\sum_{i=1}^{N(X)} \hat{W}_i N_{\hat{B}_i} + \sum_{j=1}^{M(X)} \tilde{W}_j N_{\tilde{B}_j} \mid X \leq s, N(X) > 0\right] \\ &\quad + E\left[\sum_{i=1}^{N(X)} N_{\hat{B}_i}(X + \sum_{l \neq i; l=1}^{N(X)} \hat{W}_l + \sum_{j=1}^{M(X)} \tilde{W}_j) \mid X \leq s, N(X) > 0\right] \\ &\quad + E\left[\sum_{j=1}^{M(X)} N_{\tilde{B}_j}(X + \sum_{i=1}^{N(X)} \hat{W}_i + \sum_{m \neq j; m=1}^{M(X)} \tilde{W}_m) \mid X \leq s, N(X) > 0\right]. \end{aligned} \quad (15)$$

Now from (7) we know that

$$E[N(X) \mid N(X) > 0, X] \leq 1 + \lambda P(A^c)X.$$

Hence,

$$\begin{aligned}
& E[X + \sum_{i=1}^{N(X)} \hat{W}_i + \sum_{j=1}^{M(X)} \tilde{W}_j + \sum_{i=1}^{N(X)} \hat{W}_i N_{\hat{B}_i} + \sum_{j=1}^{M(X)} \tilde{W}_j N_{\tilde{B}_j} \mid X \leq s, N(X) > 0] \\
& \leq E[X + (1 + \lambda P(A^c)X)E[W \mid A^c] + \lambda P(A)XE[W \mid A] \mid X \leq s, N(X) > 0] \\
& + E[(1 + \lambda P(A^c)X)E[WN_B \mid A^c] + \lambda P(A)XE[WN_B \mid A] \mid X \leq s, N(X) > 0] \\
& = \phi(s) (1 + \lambda E[W] + \lambda E(WN_B)) + E[W \mid A^c] + E[WN_B \mid A^c] \tag{17}
\end{aligned}$$

where  $\phi(s) = E[X \mid X \leq s, N(X) > 0]$ . Now

$$\begin{aligned}
& E[\sum_{i=1}^{N(X)} N_{\hat{B}_i} (X + \sum_{l \neq i; l=1}^{N(X)} \hat{W}_l + \sum_{j=1}^{M(X)} \tilde{W}_j) \mid X, N(X) > 0] \\
& = E[\sum_{i=1}^{N(X)} N_{\hat{B}_i} (E[X + \sum_{l \neq i; l=1}^{N(X)} \hat{W}_l + \sum_{j=1}^{M(X)} \tilde{W}_j \mid X, N(X) > 0]) \mid X, N(X) > 0].
\end{aligned}$$

This holds because for each busy period  $i$ ,  $N_{\hat{B}_i}$  is independent of  $X$ ,  $N(X)$  and  $\sum_{l \neq i; l=1}^{N(X)} \hat{W}_l + \sum_{j=1}^{M(X)} \tilde{W}_j$ . Furthermore

$$\begin{aligned}
& E[X + \sum_{l \neq i; l=1}^{N(X)} \hat{W}_l + \sum_{j=1}^{M(X)} \tilde{W}_j \mid X, N(X) > 0] \\
& \leq E[X + \sum_{l=1}^{N(X)} \hat{W}_l + \sum_{j=1}^{M(X)} \tilde{W}_j \mid X, N(X) > 0] \\
& = X + (1 + \lambda P(A^c)X)E[W \mid A^c] + \lambda P(A)XE[W \mid A] \\
& = X(1 + \lambda E(W)) + E[W \mid A^c]
\end{aligned}$$

where the last part follows from (7). We thus have

$$\begin{aligned}
& E[\sum_{i=1}^{N(X)} N_{\hat{B}_i} (X + \sum_{l \neq i; l=1}^{N(X)} \hat{W}_l + \sum_{j=1}^{M(X)} \tilde{W}_j) \mid X, N(X) > 0] \\
& \leq E[\sum_{i=1}^{N(X)} N_{\hat{B}_i} (X(1 + \lambda E(W)) + E[W \mid A^c]) \mid X, N(X) > 0] \\
& = (1 + \lambda P(A^c)X)E[N_B \mid A^c] \left( X(1 + \lambda E(W)) + E[W \mid A^c] \right) \\
& = X^2 \lambda P(A^c)E[N_B \mid A^c](1 + \lambda E(W)) + X \lambda P(A^c)E[N_B \mid A^c]E[W \mid A^c] \\
& + XE[N_B \mid A^c](1 + \lambda E(W)) + E[N_B \mid A^c]E[W \mid A^c].
\end{aligned}$$

And hence,

$$\begin{aligned}
& E\left[\sum_{i=1}^{N(X)} N_{\hat{B}_i}(X + \sum_{l \neq i; l=1}^{N(X)} \hat{W}_l + \sum_{j=1}^{M(X)} \tilde{W}_j) \mid X \leq s, N(X) > 0\right] \\
& \leq \phi'(s)\lambda P(A^c)E[N_B \mid A^c](1 + \lambda E(W)) + \phi(s)\lambda P(A^c)E[N_B \mid A^c]E[W \mid A^c] \\
& + \phi(s)E[N_B \mid A^c](1 + \lambda E(W)) + E[N_B \mid A^c]E[W \mid A^c]
\end{aligned} \tag{18}$$

where  $\phi'(s) = E[X^2 \mid X \leq s, N(X) > 0]$ . Similarly we have

$$\begin{aligned}
& E\left[\sum_{j=1}^{M(X)} N_{\tilde{B}_k}(X + \sum_{i=1}^{N(X)} \hat{W}_i + \sum_{m \neq j; m=1}^{M(X)} \tilde{W}_m) \mid X, N(X) > 0\right] \\
& \leq E\left[\sum_{j=1}^{M(X)} N_{\tilde{B}_k}(X(1 + \lambda E(W)) + E[W \mid A^c]) \mid X, N(X) > 0\right] \\
& = \lambda P(A)XE[N_B \mid A](X(1 + \lambda E(W)) + E[W \mid A^c]) \\
& = X^2\lambda P(A)E[N_B \mid A](1 + \lambda E(W)) + X\lambda P(A)E[N_B \mid A]E[W \mid A^c].
\end{aligned}$$

And hence,

$$\begin{aligned}
& E\left[\sum_{j=1}^{M(X)} N_{\tilde{B}_k}(X + \sum_{i=1}^{N(X)} \hat{W}_i + \sum_{m \neq j; m=1}^{M(X)} \tilde{W}_m) \mid X \leq s, N(X) > 0\right] \\
& \leq \phi'(s)\lambda P(A)E[N_B \mid A](1 + \lambda E(W)) \\
& + \phi(s)\lambda P(A)E[N_B \mid A]E[W \mid A^c].
\end{aligned} \tag{19}$$

Now combining equations (15), (17), (18) and (19), we have

$$\begin{aligned}
E[WN_B \mid X \leq s, A^c] & \leq \phi(s)((E[N_B \mid A^c] + 1)(1 + \lambda E[W]) + \lambda E(WN_B) \\
& + \lambda E[W \mid A^c]E(N_B)) + \phi'(s)\lambda E(N_B)(1 + \lambda E(W)) \\
& + E[WN_B \mid A^c] + E[N_B \mid A^c]E[W \mid A^c] \\
& + E[W \mid A^c].
\end{aligned} \tag{20}$$

Finally, combining equations (14) and (20), we have,

$$\begin{aligned}
E[WN_B \mid A^c] \leq & \gamma \left( E[X \mid X > s] (1 + \lambda E(W) + \lambda E(WN_B)) \right. \\
& + E[X^2 \mid X > s] \lambda (1 + \lambda E(W)) E(N_B) \Big) \\
& + (1 - \gamma) \left( E[W \mid A^c] + E[N_B \mid A^c] E[W \mid A^c] \right. \\
& + \phi(s) ((E[N_B \mid A^c] + 1)(1 + \lambda E[W]) + \lambda E(WN_B)) \\
& + \lambda E[W \mid A^c] E(N_B) + \phi'(s) \lambda E(N_B)(1 + \lambda E(W)) \\
& \left. + E[WN_B \mid A^c] \right).
\end{aligned}$$

And thus we get the upper bound:

$$\begin{aligned}
E[WN_B \mid A^c] \leq & \gamma^{-1} \left[ \gamma \left( E[X \mid X > s] (1 + \lambda E(W) + \lambda E(WN_B)) \right. \right. \\
& + E[X^2 \mid X > s] \lambda (1 + \lambda E(W)) E(N_B) \Big) \\
& + (1 - \gamma) \left( E[W \mid A^c] + E[N_B \mid A^c] E[W \mid A^c] \right. \\
& + \phi(s) ((E[N_B \mid A^c] + 1)(1 + \lambda E[W]) + \lambda E(WN_B)) \\
& + \lambda E[W \mid A^c] E(N_B) + \phi'(s) \lambda E(N_B)(1 + \lambda E(W)) \Big) \Big]. \quad (21)
\end{aligned}$$

Now from the definition of  $\gamma$ , we have

$$\begin{aligned}
E[WN_B \mid A^c] P(A^c) \leq & \frac{P(A^c)}{P(X > s)} \left[ P(X > s) \left( E[X \mid X > s] (1 + \lambda E(W) \right. \right. \\
& + \lambda E(WN_B)) + E[X^2 \mid X > s] \lambda (1 + \lambda E(W)) E(N_B) \Big) \\
& + (P(A^c) - P(X > s)) \left( E[N_B \mid A^c] E[W \mid A^c] \right. \\
& + E[W \mid A^c] + \phi(s) ((E[N_B \mid A^c] + 1)(1 + \lambda E[W]) \\
& + \lambda E(WN_B) + \lambda E[W \mid A^c] E(N_B)) \\
& \left. \left. + \phi'(s) \lambda E(N_B)(1 + \lambda E(W)) \right) \right]. \quad (22)
\end{aligned}$$

Now since  $\phi'(s) \leq E[X \mid X > s]$  and  $\phi'(s) \leq E[X^2 \mid X > s]$ , we have

$$\begin{aligned}
E[WN_B | A^c]P(A^c) &\leq \frac{P(A^c)}{P(X > s)} \left[ P(A^c) \left( E[X | X > s](1 + \lambda E(W)) \right. \right. \\
&\quad \left. \left. + \lambda E(WN_B)) + E[X^2 | X > s]\lambda(1 + \lambda E(W))E(N_B) \right) \right. \\
&\quad \left. + (P(A^c) - P(X > s)) \left( E[N_B | A^c]E[W | A^c] \right. \right. \\
&\quad \left. \left. + E[W | A^c] + \phi(s)(E[N_B | A^c](1 + \lambda E[W]) \right. \right. \\
&\quad \left. \left. + \lambda E[W | A^c]E(N_B)) \right) \right]. \tag{23}
\end{aligned}$$

Finally, using the bound on  $P(A^c)$  in (6), and ignoring the negative term we have

$$E[WN_B | A^c]P(A^c) \leq M \tag{24}$$

where

$$\begin{aligned}
M &= \frac{1}{D^2} P(X > s) \left[ E[X | X > s] \left( (1 + E[N_B | A^c])(1 + \lambda E[W]) + \lambda E(WN_B) \right. \right. \\
&\quad \left. \left. + \lambda E[W | A^c]E(N_B) \right) + E[X^2 | X > s]\lambda(1 + \lambda E(W))E(N_B) \right. \\
&\quad \left. + E[W | A^c] + E[N_B | A^c]E[W | A^c] \right]
\end{aligned}$$

and

$$D = 1 - \lambda E(X).$$

Now we have bounds for all the quantities in  $M$ . First, from the lemma, we have

$$\begin{aligned}
E[W | A^c] &\leq \frac{E[X | X > s]}{D^2} \\
E[N_B | A^c] &\leq \frac{1}{D} + \frac{\lambda E[X | X > s]}{D^2}.
\end{aligned}$$

Further, we can compute an upper bound on  $E(WN_B)$  using our decomposition :

$$\begin{aligned}
&E(WN_B) \\
&= E[X + \sum_{i=1}^{K(X)} W_i + \sum_{j=1}^{K(X)} W_j N_{B_i} + \sum_{k=1}^{K(X)} N_{B_k} (X + \sum_{j \neq k; j=1}^{K(X)} W_j)] \\
&= E(X)(1 + \lambda E(W) + \lambda E(WN_B)) + E[E[\sum_{k=1}^{K(X)} N_{B_k} (X + \sum_{j \neq k; j=1}^{K(X)} W_j) | X]] \\
&= E(X)(1 + \lambda E(W) + \lambda E(WN_B)) + E[\lambda X E(N_B)(X + (\lambda X - 1)E(W))] \\
&\leq E(X)(1 + \lambda E(W) + \lambda E(WN_B)) + \lambda(\lambda E(W) + 1)E(N_B)E(X^2).
\end{aligned}$$

We thus have

$$E(WN_B) \leq \frac{(E(X) + \lambda E(N_B)E(X^2))(\lambda E(W) + 1)}{1 - \lambda E(X)}.$$

But again since  $E(W) = \frac{E(X)}{1 - \lambda E(X)}$  and  $E(N_B) = \frac{1}{1 - \lambda E(X)}$ , we have

$$E(WN_B) \leq \frac{E(X) + \lambda E(X^2) - \lambda E(X)^2}{D^3}$$

and thus ignoring the negative term, we have

$$E(WN_B) \leq \frac{E(X) + \lambda E(X^2)}{D^3}. \quad (25)$$

We thus finally have

$$\begin{aligned} E[WN_B \mid A^c]P(A^c) &\leq M \\ &\leq \frac{P(X > s)}{D^2} \left[ E[X \mid X > s] \left( \frac{D^2 + D + \lambda E[X \mid X > s]}{D^3} \right. \right. \\ &\quad \left. \left. + \frac{\lambda E(X) + \lambda^2 E(X^2)}{D^3} + \frac{\lambda E[X \mid X > s]}{D^3} \right) \right. \\ &\quad \left. + \frac{\lambda E[X^2 \mid X > s]}{D^2} + \frac{E[X \mid X > s]}{D^2} + \frac{E[X \mid X > s]}{D^3} \right. \\ &\quad \left. + \frac{\lambda E[X \mid X > s]^2}{D^4} \right] \end{aligned} \quad (26)$$

where we have substituted all of the computed upper bounds in (24). Now by Jensen's inequality we have  $E[X \mid X > s]^2 \leq E[X^2 \mid X > s]$ . Thus we can further simplify the bound as follows:

$$\begin{aligned} E[WN_B \mid A^c]P(A^c) &\leq \frac{P(X > s)}{D^4} \left[ E[X \mid X > s] \left( 2 + D + \frac{1 + \lambda E(X) + \lambda^2 E(X^2)}{D} \right) \right. \\ &\quad \left. + E[X^2 \mid X > s] \left( \frac{2\lambda}{D} + \lambda + \frac{\lambda}{D^2} \right) \right]. \end{aligned}$$

Thus we have

$$V^s - V^* \leq E[WN_B \mid A^c]P(A^c) \leq K_1 E[X \mathbf{1}_{X>s}] + K_2 [X^2 \mathbf{1}_{X>s}] \quad (27)$$



where

$$K_1 = \frac{1}{D^4} \left( 2 + D + \frac{1 + \lambda E(X) + \lambda^2 E(X^2)}{D} \right) \text{ and} \quad (28)$$

$$K_2 = \frac{1}{D^4} \left( \frac{2\lambda}{D} + \lambda + \frac{\lambda}{D^2} \right). \quad (29)$$

Now  $\lim_{s \rightarrow \infty} E[X \mathbf{1}_{X>s}] = \lim_{s \rightarrow \infty} E[X^2 \mathbf{1}_{X>s}] = 0$  since  $E(X) < \infty$  and  $E(X^2) < \infty$ .

Thus

$$\lim_{s \rightarrow \infty} V^s - V^* = 0$$

hence proving the result. ■

### 3.1. An example

Consider  $\lambda = 1$  and assume that the service times of the jobs have the pareto distribution with the tail probability function  $1 - F(x) = P(X > x) = \frac{1}{(x+1)^3} \mathbf{1}_{\{x \geq 0\}}$ . The density function is  $f_X(x) = \frac{3}{(x+1)^4} \mathbf{1}_{\{x \geq 0\}}$ . Then we have  $E(X) = \frac{1}{2}$  and  $E[X^2] = 1$ . Thus  $D = 1 - \lambda E(X) = \frac{1}{2}$ . Then from equations (28) and (29) we find that  $K_1 = 120$  and  $K_2 = 144$ . Further we can compute

$$\begin{aligned} E[X \mathbf{1}_{\{X>s\}}] &= \frac{3s+1}{2(s+1)^3} \text{ and} \\ E[X^2 \mathbf{1}_{\{X>s\}}] &= \frac{3s^2+3s+1}{(s+1)^3}. \end{aligned}$$

Thus we have

$$V^s - V^* \leq g(s) = 120 \left( \frac{3s+1}{2(s+1)^3} \right) + 144 \left( \frac{3s^2+3s+1}{(s+1)^3} \right). \quad (30)$$

This function is plotted in fig. 2.

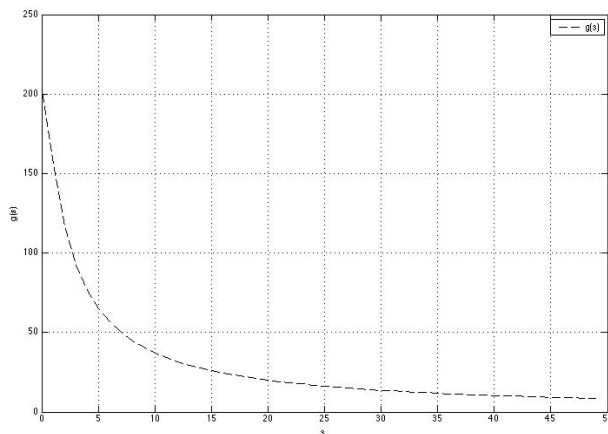


FIGURE 2: Upper bound  $g(s)$  as a function of truncation  $s$

### Acknowledgements

This work is supported by MURI grant BAA 07-036.18.

### References

- [1] Abate, J., Choudhury, G.L., Whitt, W. (1994). Waiting-time tail probabilities in queues with long-tail service-time distributions. *Queueing Systems* 16, 311-338.
- [2] Abate, J., Whitt, W. (1999). Explicit M/G/1 waiting-time distributions for a class of long-tail service-time distributions. *Operations Research Letters* 25, 25-31.
- [3] Anantharam, V. (1999). Scheduling strategies and long-range dependence. *Queueing Systems* 33, 73-89.
- [4] Asmussen, S., Kluppelberg, C. (1996). Stationary M/G/1 excursions in the presence of heavy tails. *Journal of Applied Probability* 33, 208-212.
- [5] Boxma, O.J., Dumas, V. (1998). Fluid queues with heavy-tailed activity period distributions. *Computer Communications* 21, 1509-1529.
- [6] Boxma, O.J., Cohen, J.W. (1998). The M/G/1 queue with heavy-tailed service time distribution. *IEEE Journal on Selected Areas in Communications* 16, 349-363.
- [7] Grossglauser, M., Bolot, J.-C. (1999). On the relevance of long-range dependence in network traffic. *IEEE/ACM Transactions on Networking* 7, 629-640.

- [8] Heyman, D., Lakshman, T.V. (1996). What are the implications of long-range dependence for traffic engineering? *IEEE/ACM Transactions on Networking* 4, 301-317.
- [9] Jelenkovic, P.R. (2000). On the asymptotic behavior of a fluid queue with a heavy-tailed  $M/G/1$  arrival process. Preprint, Columbia University. Submitted for publication.
- [10] Likhanov, N., Mazumdar, R.R. (2000). Loss asymptotics in large buffers fed by heterogeneous long-tailed sources. *Advances in Applied Probability* 32, 1168-1189.
- [11] Resnick, S., Samorodnitsky, G. (2000). A heavy traffic approximation for workload processes with heavy tailed service requirements. *Management Science* 46, 1236-1248.
- [12] Whitt, W. (2000). The impact of a heavy-tailed service-time distribution upon the  $M/GI/s$  waiting-time distribution. *Queueing Systems* 36, 71-87.
- [13] ZWART (2001). Queueing Systems with Heavy Tails. *PhD Thesis*, **36**, 1406–1416.